Improved imputation quality of low-frequency and rare variants in European samples using the 'Genome of The Netherlands'

Deelen, Patrick; Menelaou, Androniki; van Leeuwen, Elisabeth M.; Kanterakis, Alexandros; van Dijk, Freerk; Medina-Gomez, Carolina; Francioli, Laurent C.; Hottenga, Jouke Jan; Karssen, Lennart C.; Estrada, Karol

Published in:
European Journal of Human Genetics

DOI:
10.1038/ejhg.2014.19

Publication date:
2014

EJHG Open

European Journal of Human Genetics (2014) 22, 1321–1326

www.nature.com/ejhg

ARTICLE

Improved imputation quality of low-frequency and rare variants in European samples using the 'Genome of The Netherlands'

Patrick Deelen[1,2], Androniki Menelaou[3], Elisabeth M van Leeuwen[4], Alexandros Kanterakis[1,2], Freerk van Dijk[1,2], Carolina Medina-Gomez[5,6,7], Laurent C Francioli[3], Jouke Jan Hottenga[8], Lennart C Karssen[4], Karol Estrada[5,6,9,10], Eskil Kreiner-Møller[5,6,11], Fernando Rivadeneira[5,6,7], Jessica van Setten[3], Javier Gutierrez-Achury[1], Harm-Jan Westra[1], Lude Franke[1], David van Enckevort[2,12], Martijn Dijkstra[1,2], Heorhiy Byelas[1,2], Cornelia M van Duijn[6], Genome of the Netherlands Consortium[16], Paul I W de Bakker[3,13,14,15], Cisca Wijmenga[1] and Morris A Swertz*[,1,2]

Although genome-wide association studies (GWAS) have identified many common variants associated with complex traits, low-frequency and rare variants have not been interrogated in a comprehensive manner. Imputation from dense reference panels, such as the 1000 Genomes Project (1000G), enables testing of ungenotyped variants for association. Here we present the results of imputation using a large, new population-specific panel: the Genome of The Netherlands (GoNL). We benchmarked the performance of the 1000G and GoNL reference sets by comparing imputation genotypes with 'true' genotypes typed on ImmunoChip in three European populations (Dutch, British, and Italian). GoNL showed significant improvement in the imputation quality for rare variants (MAF 0.05–0.5%) compared with 1000G. In Dutch samples, the mean observed Pearson correlation, r^2, increased from 0.61 to 0.71. We also saw improved imputation accuracy for other European populations (in the British samples, r^2 improved from 0.58 to 0.65, and in the Italians from 0.43 to 0.47). A combined reference set comprising 1000G and GoNL improved the imputation of rare variants even further. The Italian samples benefitted the most from this combined reference (the mean r^2 increased from 0.47 to 0.50). We conclude that the creation of a large population-specific reference is advantageous for imputing rare variants and that a combined reference panel across multiple populations yields the best imputation results.
European Journal of Human Genetics (2014) **22,** 1321–1326; doi:10.1038/ejhg.2014.19; published online 4 June 2014

Keywords: genotype imputation; GWAS; GoNL; rare variants; reference sets; reference panel

INTRODUCTION

Although genome-wide association studies (GWAS) have been very effective in identifying loci associated with diseases or traits,[1] it has proved difficult to fine-map the association signals to causal variants.[2,3] To overcome these limitations, there has been increasing interest in the interrogation of less frequent variants, especially given the enrichment of deleterious alleles at low frequencies.[4–7] There are specialized chips that can assess a larger number of rare variants, like the ImmunoChip[8] or Metabochip,[9] although they do not provide uniform genome-wide coverage. Hence, most investigators will use statistical imputation from SNP arrays in GWAS using dense reference panels.

Imputation using a densely typed reference set can be performed to infer untyped variants that can be used to improve the power of a GWAS,[10] and there are numerous examples in which imputation has effectively enriched the results in GWAS.[11,12] Although most large studies have so far been based on meta-analysis of HapMap-based imputations across cohorts, the primary limitation is that HapMap is essentially restricted to common variation (MAF > 5%). Thanks to the sequencing of larger samples, such as 1000G, more complete reference panels are now being assembled, setting off a new wave of meta-analyses.

The power of detecting an association in a GWAS is determined by its sample size and effective genome-wide coverage of the included

[1]University of Groningen, University Medical Center Groningen, Department of Genetics, Groningen, The Netherlands; [2]University of Groningen, University Medical Center Groningen, Genomics Coordination Center, Groningen, The Netherlands; [3]Department of Medical Genetics, University Medical Center Utrecht, Utrecht, The Netherlands; [4]Department of Epidemiology, Genetic Epidemiology Unit, Erasmus Medical Center, Rotterdam, The Netherlands; [5]Department of Internal Medicine, Erasmus Medical Center, Rotterdam, The Netherlands; [6]Department of Epidemiology, Erasmus University Medical Center, Rotterdam, The Netherlands; [7]Netherlands Genomics Initiative (NGI)-sponsored Netherlands Consortium for Healthy Aging (NCHA), Rotterdam, The Netherlands; [8]Department of Biological Psychology, VU University Amsterdam, Amsterdam, The Netherlands; [9]Department of Medicine, Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA; [10]Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA; [11]COPSAC; Copenhagen Prospective Studies on Asthma in Childhood; Faculty of Health Sciences, University of Copenhagen, Copenhagen, Denmark; [12]NBIC BioAssist, Netherlands Bioinformatics Center, Nijmegen, The Netherlands; [13]Department of Epidemiology, University Medical Center Utrecht, Utrecht, The Netherlands; [14]Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA; [15]Broad Institute of Harvard and MIT, Cambridge, MA, USA [16]Genome of the Netherlands Consortium members are listed before the references.
*Correspondence: Dr MA Swertz, Genomics Coordination Center, University Medical Center Groningen, University of Groningen, Groningen 9700 RB, The Netherlands Tel: +31 6 52 60 65 01; Fax: + 31 50 361 7230; E-mail: m.a.swertz@gmail.com
Received 13 August 2013; revised 1 November 2013; accepted 16 January 2014; published online 4 June 2014

variants, among other things.[13,14] The effective coverage depends directly on the number and quality of the imputed genotypes.[15] In turn, the quality of the reference panel will depend largely on the number of samples, the quality of the haplotypes, and the number of variants included.[16]

The Genome of The Netherlands (GoNL) has the potential to provide a good imputation reference panel. GoNL is a population-based sequencing project, in which 769 Dutch samples were sequenced at, on average, 14× coverage.[17] In particular, the fact that GoNL sequenced trios (231) or quartets (19) has enabled improved haplotype phasing by using one of the children.[18] The GoNL imputation reference set contains 998 unrelated haplotypes. In this paper, we report a quantitative analysis to assess the quality of imputed genotypes from using both GoNL and 1000G in Dutch and other European populations.

We adopted a 'gold standard' approach using samples genotyped on two distinct platforms, HumanHap550 and ImmunoChip. Hap550 is a commonly used genotyping chip designed to tag as many haplotypes as possible using common variants. ImmunoChip, however, is a fine-mapping chip: it contains a large number of low-frequency and rare variants for a limited number of loci (primarily selected on the basis of loci identified in immune-related traits). Starting from the Hap550-genotyped SNPs, we were able to impute a large number of variants present on ImmunoChip. We then compared these imputed genotypes with the measured ('gold standard') genotypes on ImmunoChip to quantify the imputation performance. We have such a data set for three European populations: the Dutch, British, and Italians. For each population we used 745 samples genotyped on both platforms. These three populations allowed us to ascertain population-specific differences in the imputation quality of SNPs.

MATERIALS AND METHODS

Genome of the Netherlands

GoNL is a project in which 769 individuals from different Dutch provinces were sequenced at, on average, 14× coverage.[17] All samples are part of either one of the 231 trios or one of the 19 quartets. The phasing was performed using the trio information,[18] and for the quartets one of the children was used to enhance the phasing. Because of sequence failures of two parents, from different trios, these samples were excluded from the imputation reference set. Instead, from these two trios, we used the haplotype of the child that was not present in the other parent. This resulted in an imputation reference set containing 998 unrelated haplotypes. We used GoNL release 4 for all our analyses (see http://www.nlgenome.nl). The current GoNL release 5 also contains over one million indels but did not change the SNPs.

Benchmarking samples

Samples from a celiac disease patient cohort were selected, as they had been genotyped on both the Hap550 and ImmunoChip.[19] The 745 Dutch and the 745 British samples were all cases, whereas the 745 Italian samples comprised 371 cases and 374 controls. The clustering for the genotype calling of the ImmunoChip data was performed manually in the past, to ensure proper genotyping results.

The Hap550 (516 426 SNPs) data were filtered on MAF > 1% and HWE *P*-value > 1E-4 for each population separately. The ImmunoChip (113 991 SNPs) data were filtered on MAF > 0.05% and HWE *P*-value of 1E-4. Both data sets are filtered on variants present in both the 1000G reference set and the GoNL reference set. After QC the Dutch, British, and Italian Hap550 data contain 509 888, 509 984, and 510 225 SNPs, respectively. The ImmunoChip data contain in the same order 107 383, 107 212, and 107 611 SNPs.

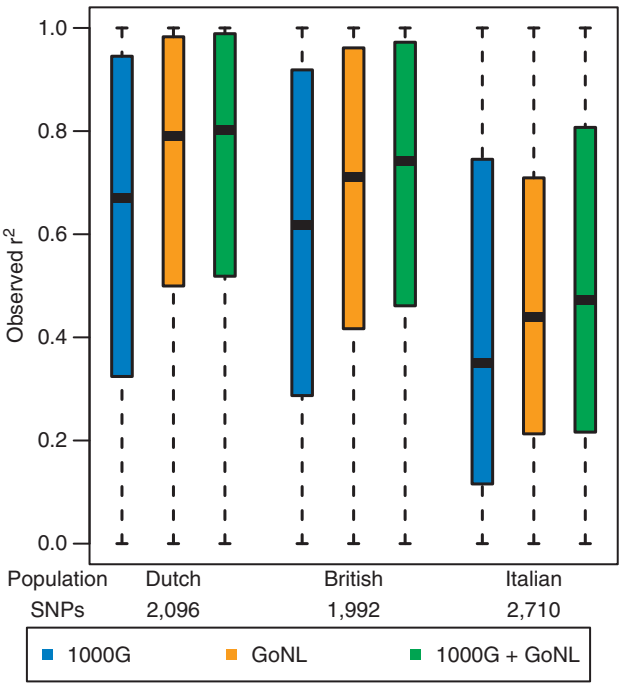

Figure 1 Comparison of imputation quality of rare variants using the 1000G data, GoNL, and the combined reference panel.

Combining 1000G and GoNL data

The reference set combining data from 1000G and GoNL was created using the Impute2 option: '--merge_ref_panels'. This merged reference set was written to a file and subsequently used for the benchmarking. As our benchmarking data are filtered for variants present in both reference sets, we did not assess the imputations of variants that are unique to either reference set.

Pre-phasing

The 745 samples for each population were pre-phased using SHAPEIT2.[15] This was done per chromosome using the default settings.

Imputation

The imputations were performed using Impute2 2.3.0.[16] The different populations were imputed separately and in chunks of 5 Mb. For the comparison using an equal number of identical European haplotypes, we performed an imputation using all 379 European 1000G samples and a random selection of 379 GoNL samples. The random selection of GoNL samples was performed stratified on the Dutch provinces. These samples were selected using the Impute2 option: '--exclude_samples_h'.

We used MOLGENIS compute[20] to implement the imputation pipeline, run the 8835 imputation chunks in parallel on a PBS compute cluster, and keep track of the 15 imputations (five for each population). All pipelines are available as open source via http://www.molgenis.org/wiki/ComputeStart.

Gold standard method

As stated above, we used samples genotyped on two distinct platforms. We imputed the Hap550 genotypes from these samples and compared the imputed genotypes with the SNPs previously present only in the ImmunoChip data. We used the ImmunoChip data as our 'gold standard'. The concordance between imputed genotypes and ImmunoChip genotypes was determined by calculating the Pearson correlation r^2 between the imputed dosage and ImmunoChip-observed genotypes. The mean concordances were calculated for three MAF bins: rare (≥0.05% and <0.5%), low-frequency (≥0.5% and <5%), and common (>5%) SNPs. The MAF used to stratify the SNPs into the bins was calculated separately for each population. The results were plotted using R

2.14.2.[21] The significance of the differences between the reference sets was calculated using the Wilcoxon signed-rank test implementation in R.

Principal component analysis

The principal component analysis was performed using the EIGENSOFT 4.2 package.[22] The components were calculated using the European 1000G, GoNL, and the 3 GWAS data sets that we used for benchmarking. Before the components were calculated, all data sets were filtered to include only variants with MAF > 5%. A joint data set, featuring variants present in all five data sets, was created. This data set was again filtered for MAF > 5%; the merged data were also filtered on HWE > 1E-4 and a call rate of 95%. This data set was pruned using PLINK 1.07[23] with the '–indep-pairwise' option, windows: 1000, step: 5, r^2 threshold: 0.2. The first component explained 0.33% of the variation and the second 0.10%. All subsequent components described less than 0.06%.

RESULTS

We stratified our analysis into three groups: common variants (MAF ≥ 5%), low-frequency variants (MAF 0.5–5%), and rare variants (MAF 0.05–0.5%). We focused mainly on the rare variants, as these are more difficult to impute and most can be gained in terms of imputation quality when using a better reference set. We observed a large increase in the imputation quality of rare variants when using GoNL as the reference compared with 1000G (Figure 1, Table 1). The mean observed Pearson correlation (r^2) showed a significant increase from 0.61 to 0.71 for Dutch samples (Wilcoxon *P*-value = 7.16E-60). The British and Italian imputations also showed a significant improvement when imputing rare variants, from 0.58 to 0.65 ($P = 3.70E-35$) and from 0.43 to 0.47 ($P = 2.64E-13$), respectively. GoNL also significantly outperformed the 1000G reference set in the imputation of variants with higher MAFs (Supplementary Figures/Supplementary Appendices S1, S2, S3).

Using a combined reference set composed of the 1000G and GoNL samples, we could improve the imputation further. The imputation of rare variants using the combined reference in Dutch and British samples showed a small increase in quality compared with GoNL-only imputation (0.02 ($P = 1.16E-03$) and 0.02 ($P = 2.70E-05$), respectively). The Italians benefitted most from the combined reference with an increase of 0.04 ($P = 3.62E-30$) compared with a GoNL-only reference, resulting in a mean concordance for rare variants of 0.5. The differences in imputation quality when using the combined reference set for more frequent alleles were either very small or not significant (Supplementary Figure S1, Supplementary Tables S2 and S3).

A striking trend in these results is that the imputation quality of rare variants in the Italian samples is lower than that in Dutch and British samples. The Dutch and Italian samples were genotyped at the same center and have similar call rates, and there were no indications that the genotyping quality of the Italian samples was lower. However, a principal component analysis revealed that the Italian samples were

Table 1 Mean observed r^2 of rare variants

Reference set	*Dutch*	*British*	*Italian*
1000G	0.61	0.58	0.43
GoNL	0.71	0.65	0.47
1000G + GoNL	0.72	0.67	0.50

Abbreviation: GoNL, The Genome of The Netherlands.
Differences in the mean imputation quality between the reference sets was significant for each population ($P < 0.001$).

Table 2 Mean observed r^2 of rare variants for reference sets of equal sample size from 1000G and GoNL (all of European descent)

Reference set	*Dutch*	*British*	*Italian*
1000G European	0.59	0.57	0.40
GoNL random subset 379 samples	0.68	0.64	0.45

Abbreviation: GoNL, The Genome of The Netherlands.
Differences in the mean imputation quality between the reference sets of equal sample size was significant for each population ($P < 0.001$).

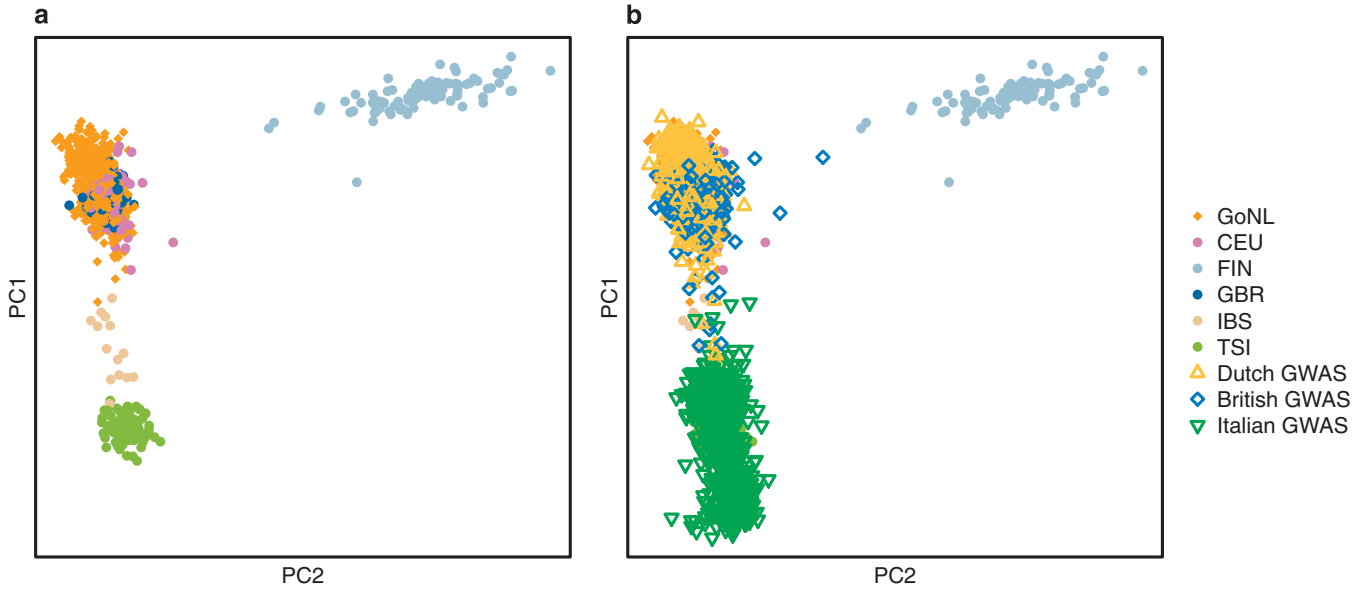

Figure 2 Clustering of reference and study samples. PC1 and PC2 reveal three main clusters: Tuscans from Italy (TSI), Finnish (FIN), and a Western European cluster with the CEU (Utah Residents with Northern and Western European ancestry), the GBR (British) and the GoNL samples (**a**). **b** shows that most of our GWAS samples clustered in a similar way to the corresponding 1000G/GoNL samples.

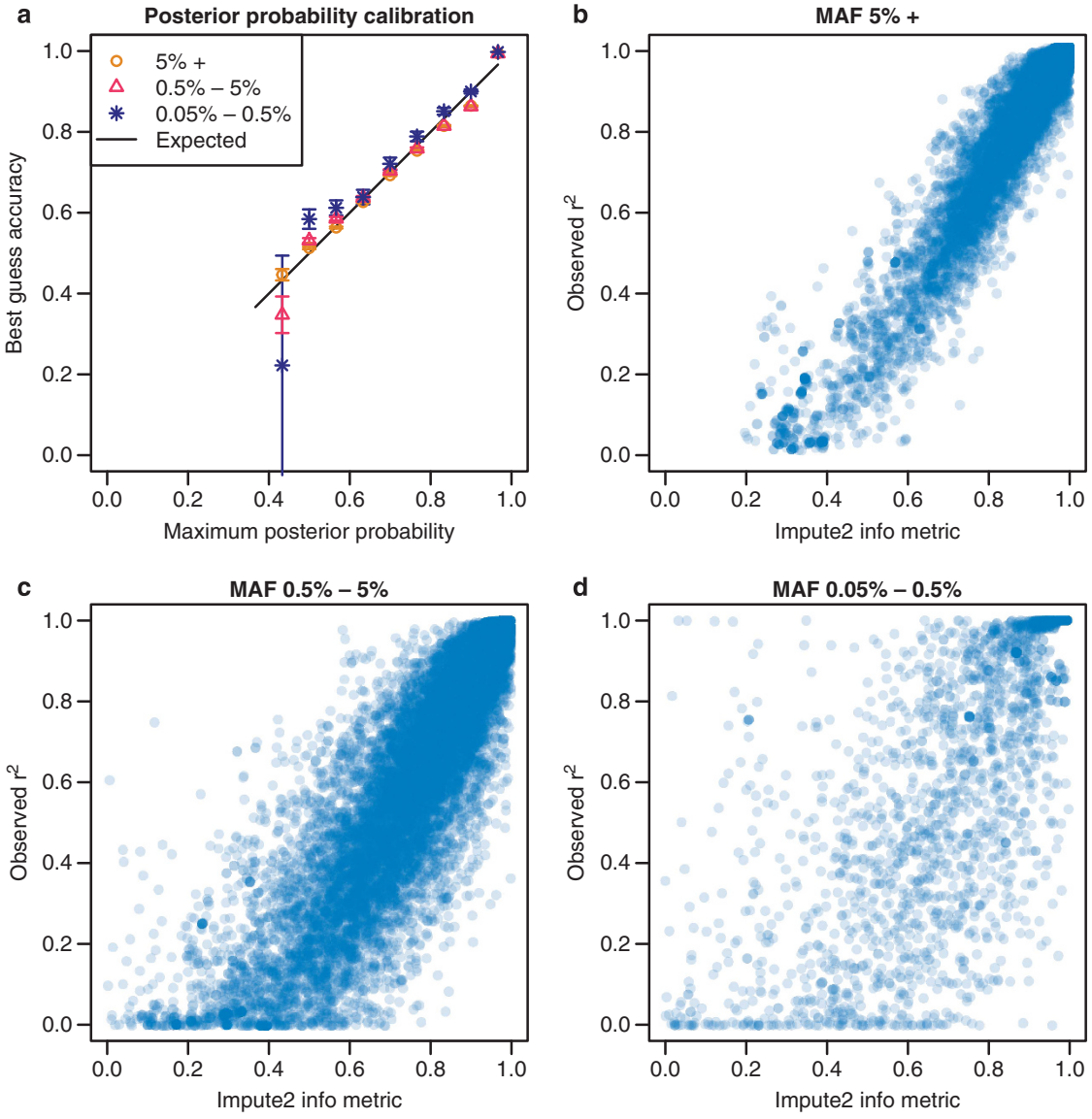

Figure 3 Calibration of posterior probabilities. The posterior probabilities were, in general, well calibrated, although there were a few deviations from the expected accuracy (**a**). For common and low-frequency variants (**b** and **c**), we observed a strong correlation (r^2 0.97 and 0.91, respectively) between the impute2 info metric and the observed r^2. However, for the rare variants (**d**), the relation between predicted and observed quality was less profound. We also observed a correlation of 0.70 and several large deviations from the diagonal.

not as well represented by either 1000G or GoNL compared with the Dutch and British GWAS samples used for benchmarking (Figure 2).

We assessed whether the better performance of GoNL compared with 1000G was due to the larger number of European haplotypes in the reference set (998 *vs.* 758 in 1000G). We did this by performing an imputation using solely the 379 European samples in 1000G and a random subset of 379 GoNL samples. We found that the GoNL subset also significantly outperformed the European 1000G subset (Table 2).

Our experimental design also allowed us to assess the calibration of the posterior probabilities of the genotypes as they are output by Impute2. We observed that the posterior probabilities were, in general, well calibrated, although we did observe a few deviations for low-frequency and rare variants (Figure 3a). To ascertain whether these deviations in posterior probabilities affect the predicted imputation quality, the Impute2 info metric, we plotted the predicted quality against the observed r^2. This showed a strong correlation between the predicted and observed quality for common variants and low-frequency variants (correlation of 0.97 and 0.91, respectively; Figures 3b and c). However, the info metric is not as accurate for rare variants, and the correlation with the observed r^2 dropped to 0.70 (Figure 3d). We also observed some discrepancies wherein a near-perfect imputation was predicted while in fact there was poor imputation, and vice versa when assessing rare variants.

DISCUSSION

We have shown that the new GoNL reference set provides higher downstream imputation accuracy than the 1000G reference set, not only for Dutch samples but also for other European populations studied in this paper. Aside from the increase in the imputation quality of rare variants in Dutch samples from 0.61 (1000G) to 0.71 (GoNL), we also observed an increase in imputation quality in British

(0.58–0.65) and Italian (0.43–0.47) samples. We show that GoNL yielded better imputed genotypes for at least these European populations. A combined reference set, of 1000G and GoNL, increased the mean imputation quality of rare variants even further to 0.72, 0.67, and 0.50 for the Dutch, British and Italians, respectively.

By selecting an identical number of European haplotypes from 1000G and GoNL, we showed a strong added value for GoNL in all the tested populations, confirming that the trio design of GoNL and the resultant accurate haplotypes aid the downstream imputation quality. We also observed a population-specific added value of GoNL when imputing Dutch samples. The added value (ie mean increase in imputation quality) was largest when comparing GoNL with 1000G in imputing the Dutch samples. Of course, it was already known that a better matched reference set will result in better imputed genotypes;[13] however, the results from this paper were based on low-frequency variants and we show that there is also an inter-European effect of reference sets.

It is important to note that we only assessed variants present on the ImmunoChip. Although these variants were not randomly selected, we have no reason to assume that the imputation quality will be positively biased or that they do not represent low-frequency variants in general. The ImmunoChip was made to fine-map loci previously associated with autoimmune diseases using a large number of low-frequency and rare variants.

We were encouraged by the observation that the posterior probabilities were, in general, well calibrated with respect to the gold standard genotypes. We observed no adverse effects on the accuracy of the Impute2 info metrics, although for rare variants we did observe a few instances with large deviations between the predicted and observed quality. This is in line with previous observations.[24] This observed inaccuracy also emphasizes the importance of validating associations from imputed genotypes.

It was shown earlier that a larger and more diverse reference set can improve the imputation of low-frequency variants.[25] We observed that a combination of 1000G and GoNL showed limited added value for the imputation of rare variants in the Dutch and British samples. It was, however, interesting to observe that the imputation of the Italian samples was improved more by this combined reference panel, leading us to speculate that populations that are poorly represented in the reference panel benefit more from a large and diverse reference set. Despite the limited added value for the Dutch and British data sets, such a large reference set may still be of interest for consortia aiming to impute cohorts of both European and non-European origin. All these cohorts can be imputed using the same combined reference set and then use Impute2 to automatically select the best matching haplotypes.[26] We should note that we were only able to assess variants present in both reference sets, as there are very few variants on the ImmunoChip that are unique to either GoNL or 1000G. Nonetheless, our results show that population-specific reference sets and cosmopolitan panels, such as 1000G, can augment each other. This even holds true for the imputation of samples with ancestry other than those present in the population-specific reference sets, which provides further motivation for international efforts towards large and integrated reference sets.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This study was made possible by rainbow grant 2 from BBMRI-NL to MS, a research infrastructure financed by the Netherlands Organization for Scientific Research (NWO project 184.021.007). We thank the Target project (http://www.rug.nl/target) for providing the compute infrastructure, and the BigGrid/eBioGrid project (http://www.ebiogrid.nl) for sponsoring the pipeline implementation. We thank Jackie Senior for careful reading and editing the manuscript. This study made use of data generated by the 'Genome of the Netherlands' project, which is funded by the Netherlands Organization for Scientific Research (grant no. 184021007). The data were made available as a Rainbow Project of BBMRI-NL. Samples were contributed by LifeLines (http://lifelines.nl/lifelines-research/general), the Leiden Longevity Study (http://www.healthy-ageing.nl; http://www.langleven.net), the Netherlands Twin Registry (NTR: http://www.tweelingenregister.org), the Rotterdam studies, (http://www.erasmus-epidemiology.nl/rotterdamstudy), and the Genetic Research in Isolated Populations program (http://www.epib.nl/research/geneticepi/research.html#gip). The sequencing was carried out in collaboration with BGI (Shenzhen, China).

AUTHOR CONTRIBUTIONS

PD, AM, MAS, PIWdB, and CW wrote the main manuscript.
All the authors contributed to the discussion of experimental design in 'Genome of The Netherlands' imputation working group. EMvL, AK, LCK, CM-G, JjH, and FvD revised the manuscript. PD, FvD, MD, HB, LCF, H-JW, AK, EK-M, and CM-G contributed to the implementation of the analysis.

GENOME OF THE NETHERLANDS CONSORTIUM

Analysis group: Morris A. Swertz[6,7] (Co-Chair), Laurent C. Francioli[1], Freerk van Dijk[6,7], Androniki Menelaou[1], Pieter B.T. Neerincx[6,7], Sara L. Pulit[1], Patrick Deelen[6,7], Clara C. Elbers[1], Pier Francesco Palamara[2], Itsik Pe'er[2,8], Abdel Abdellaoui[9], Wigard P. Kloosterman[1], Mannis van Oven[10], Martijn Vermaat[11], Mingkun Li[12], Jeroen F.J. Laros[11], Mark Stoneking[12], Peter de Knijff[13], Manfred Kayser[10], Jan H. Veldink[14], Leonard H. van den Berg[14], Heorhiy Byelas[6,7], Johan T. den Dunnen[11], Martijn Dijkstra[6,7], Najaf Amin[15], K. Joeri van der Velde[6,7], Jouke Jan Hottenga[9], Jessica van Setten[1], Elisabeth M. van Leeuwen[15], Alexandros Kanterakis[6,7], Mathijs Kattenberg[9], Lennart C. Karssen[15], Barbera D.C. van Schaik[16], Jan Bot[17], Isaäc J. Nijman[1], David van Enckevort[18], Hailiang Mei[18], Vyacheslav Koval[19], Kai Ye[20,21], Eric-Wubbo Lameijer[21], Matthijs H. Moed[21], Jayne Y. Hehir-Kwa[22], Robert E. Handsaker[5,23], Shamil R. Sunyaev[4,5], Mashaal Sohail[4,5], Fereydoun Hormozdiari[24], Tobias Marschall[25], Alexander Schönhuth[25], Victor Guryev[26], Paul I.W. de Bakker[1,3-5] (Co-Chair);

Cohort collection and sample management group: P. Eline Slagboom[21], Marian Beekman[21], Anton J.M. de Craen[21], H. Eka D. Suchiman[21], Albert Hofman[15], Cornelia van Duijn[15], Dorret I. Boomsma[9], Gonneke Willemsen[9], Bruce H. Wolffenbuttel[27], Mathieu Platteel[6], Steven J. Pitts[28], Shobha Potluri[28], David R. Cox[28,34];

Whole-genome sequencing: Qibin Li[29], Yingrui Li[29], Yuanping Du[29], Ruoyan Chen[29], Hongzhi Cao[29], Ning Li[30], Sujie Cao[30], Jun Wang[29,31,32];

Ethical, Legal, and Social Issues: Jasper A. Bovenberg[33]

Steering committee: Cisca Wijmenga[6,7] (Principal Investigator), Morris A. Swertz[6,7], Cornelia M. van Duijn[15], Dorret I. Boomsma[9], P. Eline Slagboom[21], Gertjan B. van Ommen[11], Paul I.W. de Bakker[1,3-5]

1: Department of Medical Genetics, University Medical Center Utrecht, Utrecht, The Netherlands

2: Department of Computer Science, Columbia University, New York, NY, USA

3: Department of Epidemiology, University Medical Center Utrecht, Utrecht, The Netherlands

4: Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

5: Broad Institute of Harvard and MIT, Cambridge, MA, USA

6: Department of Genetics, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands

7: Genomics Coordination Center, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands

8: Department of Systems Biology, Columbia University, New York, NY, USA

9: Department of Biological Psychology, VU University Amsterdam, Amsterdam, The Netherlands

10: Department of Forensic Molecular Biology, Erasmus Medical Center, Rotterdam, The Netherlands

11: Leiden Genome Technology Center, Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands

12: Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

13: Forensic Laboratory for DNA Research, Department of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands

14: Department of Neurology, University Medical Center Utrecht, Utrecht, The Netherlands

15: Department of Epidemiology, Erasmus Medical Center, Rotterdam, The Netherlands

16: Bioinformatics Laboratory, Department of Clinical Epidemiology, Biostatistics and Bioinformatics, Amsterdam Medical Center, Amsterdam, The Netherlands

17: SURFsara, Science Park, Amsterdam, The Netherlands

18: Netherlands Bioinformatics Centre, Nijmegen, The Netherlands

19: Department of Internal Medicine, Erasmus Medical Center, Rotterdam, The Netherlands

20: The Genome Institute, Washington University, St. Louis, MO, USA

21: Section of Molecular Epidemiology, Department of Medical Statistics and Bioinformatics, Leiden University Medical Center, Leiden, The Netherlands

22: Department of Human Genetics, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands

23: Department of Genetics, Harvard Medical School, Boston, MA, USA

24: Department of Genome Sciences, University of Washington, Seattle, WA, USA

25: Centrum voor Wiskunde en Informatica, Life Sciences Group, Amsterdam, The Netherlands

26: European Research Institute for the Biology of Ageing, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands

27: Department of Endocrinology, University Medical Center Groningen, Groningen, The Netherlands

28: Rinat-Pfizer Inc, South San Francisco, CA, USA

29: BGI-Shenzhen, Shenzhen, China

30: BGI-Europe, Copenhagen, Denmark

31: Department of Biology, University of Copenhagen, Copenhagen, Denmark

32: The Novo Nordisk Foundation Center for Basic Metabolic Research, University of Copenhagen, Copenhagen, Denmark

33: Legal Pathways Institute for Health and Bio Law, Aerdenhout, The Netherlands

34: Deceased

1 Hindorff LA, Sethupathy P, Junkins HA *et al*: Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA* 2009; **106**: 9362–9367.
2 Maller JB, McVean G, Byrnes J *et al*: Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat Genet* 2012; **44**: 1294–1301.
3 Shea J, Agarwala V, Philippakis AA *et al*: Comparing strategies to fine-map the association of common SNPs at chromosome 9p21 with type 2 diabetes and myocardial infarction. *Nat Genet* 2011; **43**: 801–805.
4 Kryukov GV, Pennacchio LA, Sunyaev SR: Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am J Hum Genet* 2007; **80**: 727–739.
5 Cirulli ET, Goldstein DB: Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet* 2010; **11**: 415–425.
6 Lee S, Wu MC, Lin X: Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* 2012; **13**: 762–775.
7 Huyghe JRJ, Jackson AUA, Fogarty MMP *et al*: Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. *Nat Genet* 2013; **45**: 197–201.
8 Cortes A, Brown MA: Promise and pitfalls of the Immunochip. *Arthritis Res Ther* 2011; **13**: 101.
9 Keating BJ, Tischfield S, Murray SS *et al*: Concept, design and implementation of a cardiovascular gene-centric 50 k SNP array for large-scale genomic association studies. *PLoS One* 2008; **3**: e3583.
10 Hao K, Chudin E, McElwee J, Schadt E: Accuracy of genome-wide imputation of untyped markers and impacts on statistical power for association studies. *BMC Genet* 2009; **10**: 27.
11 Holm H, Gudbjartsson DF, Sulem P *et al*: A rare variant in MYH6 is associated with high risk of sick sinus syndrome. *Nat Genet* 2011; **43**: 316–320.
12 Li Y, Willer C, Sanna S, Abecasis G: Genotype imputation. *Annu Rev Genomics Hum Genet* 2009; **10**: 387–406.
13 De Bakker PIW, Yelensky R, Pe'er I *et al*: Efficiency and power in genetic association studies. *Nat Genet* 2005; **37**: 1217–1223.
14 Flannick J, Korn JM, Fontanillas P *et al*: Efficiency and power as a function of sequence coverage, SNP array density, and imputation. *PLoS Comput Biol* 2012; **8**: e1002604.
15 Zheng J, Li Y, Abecasis G, Scheet P: A comparison of approaches to account for uncertainty in analysis of imputed genotypes. *Genet Epidemiol* 2011; **35**: 102–110.
16 Howie B, Donnelly P, Marchini J: A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* 2009; **5**: e1000529.
17 Boomsma DI, Wijmenga C, Slagboom EP *et al*: The Genome of the Netherlands: design, and project goals. *Eur J Hum Genet* 2014; **22**: 221–227.
18 Menelaou A, Marchini J: Genotype calling and phasing using next-generation sequencing reads and a haplotype scaffold. *Bioinformatics* 2013; **29**: 84–91.
19 Trynka G, Hunt KA, Bockett NA *et al*: Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat Genet* 2011; **43**: 1193–1201.
20 Byelas H, Dijkstra M, Neerincx P *et al*: Scaling bio-analyses from computational clusters to grids. *Proceedings of the 5th International Workshop on Science Gateways (IWSG 2013)*. CEUR-WS.org; Zurich, Switzerland. ISSN: 1613–0073.
21 R Core Team: *R: A language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2008; p409.
22 Price AL, Patterson NJ, Plenge RM *et al*: Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006; **38**: 904–909.
23 Purcell S, Neale B, Todd-Brown K *et al*: PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 2007; **81**: 559–575.
24 Li L, Li Y, Browning SR *et al*: Performance of genotype imputation for rare variants identified in exons and flanking regions of genes. *PLoS One* 2011; **6**: e24945.
25 Jostins L, Morley K, Barrett J: Imputation of low-frequency variants using the HapMap3 benefits from large, diverse reference sets. *Eur J Hum Genet* 2011; **19**: 662–666.
26 Howie B, Marchini J, Stephens M: Genotype imputation with thousands of genomes. *G3 genes-genomes-Genet* 2011; **1**: 457–470.

Supplementary Information accompanies this paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)